Robust Design for the Estimation of a Threshold Probability

Rui Hu

MacEwan University

August 09, 2017

Latest Advances in the Theory and Applications of Design and Analysis of Experiments

Banff International Research Station

Model for the Observed Data

- Let $\mathbf{t} = (t_1, ..., t_d)^T$, $d \geq 1$, be a 'location' from a set $\mathcal{T} = \{\mathbf{t}_1, \cdots, \mathbf{t}_N\} \subset \mathbb{R}^d$.
- Observations $\mathbf{y}_n = (y(\mathbf{t}_{s1}), \cdots, y(\mathbf{t}_{sn}))^T$ will be obtained once sample locations $\mathcal{S} = \{\mathbf{t}_{s1}, \cdots, \mathbf{t}_{sn}\} \subset \mathcal{T}$ are chosen.
- A model used to describe the observed data is

$$Y(\mathbf{t}) = \mu(\mathbf{t}) + \varepsilon(\mathbf{t})$$

  - $\mu(\mathbf{t}) = \eta(\mathbf{t}) + \delta(\mathbf{t})$, $\eta(\mathbf{t})$ is the deterministic mean perturbed by stochastic errors $\delta(\mathbf{t})$, and $\varepsilon(\mathbf{t})$ is uncorrelated, additive measurement error.
  - $\{\delta(\mathbf{t}) : \mathbf{t} \in \mathcal{T}\}$ and $\{\varepsilon(\mathbf{t}) : \mathbf{t} \in \mathcal{T}\}$ are independent.

Threshold Probability

- Consider the probability that the value of the μ process, at a given location $\mathbf{t}$, is above a fixed threshold u_*, i.e.

$$z(\mathbf{t}) = P\left(\mu(\mathbf{t}) > u_*\right).$$

- The probability $z(\mathbf{t})$ is called the ‘threshold probability’.

- A natural and optimal (Cressie 1993) estimator of $z(\mathbf{t})$ is

$$\hat{z}_n(\mathbf{t}) := E\left[\mathbf{1}(\mu(\mathbf{t}) > u_*)|\mathbf{y}_n\right].$$

Further Assumptions

- $\delta(\mathbf{t})$ and $\varepsilon(\mathbf{t})$ are Gaussian processes.
- $\{\delta(\mathbf{t})|\mathbf{t} \in \mathcal{T}\}$ has covariance matrix $\mathbf{G}_{N\times N} = (g(\mathbf{t}_i, \mathbf{t}_j))_{i,j=1}^{N}$ and $\{\varepsilon(\mathbf{t})|\mathbf{t} \in \mathcal{T}\}$ has uniformly bounded variance matrix $\mathbf{H}_{N\times N} = diag(h(\mathbf{t}_i))_{i=1}^{N}$, with $h(\mathbf{t}) \in [h_1, h_N]$ for all $\mathbf{t} \in \mathcal{T}$ and $0 < h_1 \leq h_N < \infty$.

Lemma

(Santner, Williams and Notz 2003) Assume that $E(\mu(\mathbf{t})) = \mathbf{f}^T(\mathbf{t})\theta$ *for a* $p-$*dimensional vector of functions* $\mathbf{f}(\mathbf{t}) = (f_1(\mathbf{t}), \cdots, f_p(\mathbf{t}))^T$*, and that* $\theta \sim \mathcal{U}_{\mathbb{R}^p}$*, the improper uniform distribution over* $\mathbb{R}^p$*. Define* $\mathbf{F}_n = (\mathbf{f}(\mathbf{t}_{s1}), \cdots, \mathbf{f}(\mathbf{t}_{sn}))^T$*. Then*

$$\mu(\mathbf{t})|\mathbf{y}_n \sim GP(\hat{\mu}(\mathbf{t}), \sigma^2_{n\mathbf{t}}(\mathbf{G})),$$

where the conditional (given data $\mathbf{y}_n$*) mean of* $\mu(\mathbf{t})$ *is the best linear unbiased predictor (BLUP)*

$$\hat{\mu}(\mathbf{t}) = \mathbf{a}^T_{n\mathbf{t}}(\mathbf{G})\mathbf{y}_n.$$

The conditional variance is the mean squared prediction error of the BLUP.

Based on Lemma 1

$$\hat{z}_n(\mathbf{t}) = \Phi\left(\frac{\hat{\mu}(\mathbf{t}) - u_*}{\sigma_{n\mathbf{t}}(\mathbf{G})}\right).$$

- The estimator $\hat{z}_n(\mathbf{t})$ is derived under quite restrictive assumptions:
  - **(U1)** the covariance structure of $\delta(\mathbf{t})$ is known.
  - **(U2)** The mean $E(\mu(\mathbf{t}))$ is linear in these regressors $\mathbf{f}(\mathbf{t})$.

- The investigator may not address these challenges at the estimation/prediction stage, but hopes to do so through the design.
- The estimate $\hat{z}_n(\mathbf{t})$ is still computed based on a nominal covariance matrix ($\mathbf{U}_1$) and the possibly incorrect response ($\mathbf{U}_2$).

- Our interest is to develop a design such that the estimate of $z(\mathbf{t})$ is robust against uncertainties (**U1**) and (**U2**).

- To address (**U1**) we assume that the covariance matrix $\mathbf{G}_{N\times N}$ varies over a neighbourhood $\mathcal{G}$

$$\mathcal{G} = \left\{ \begin{array}{c} \mathbf{G_d}\colon \mathbf{G_d} = \mathbf{U}\text{diag}(\lambda_i e^{\mathbf{d}/\sqrt{n}})_{i=1}^{N}\mathbf{U}^T, \\ \mathbf{d} \text{ is a bounded random variable with} \\ \text{mean } 0 \text{ and standard deviation } \omega_{\mathbf{d}}^2\,, \\ \text{and } -\infty < d_1 \le \mathbf{d} \le d_2 < \infty \end{array} \right\}.$$

Here $\lambda_1 \le ... \le \lambda_N$ are the eigenvalues of a nominal, positive definite covariance matrix and $\mathbf{U}$ is the orthogonal matrix whose columns are the corresponding eigenvectors.

- To address (**U2**) suppose that $E(Y(\mathbf{t})) \approx \mathbf{f}^T(\mathbf{t})\theta$.
- Define $\psi(\mathbf{t})$ as

$$E(Y(\mathbf{t})) = \mathbf{f}^T(\mathbf{t})\theta + \frac{\psi(\mathbf{t})}{\sqrt{n}}.$$

- The parameter vector θ is defined as

$$\theta = \arg\min_{\mathbf{v}} \sum_{\mathbf{t}\in\mathcal{T}} \left(E(Y(\mathbf{t})) - \mathbf{f}^T(\mathbf{t})\mathbf{v}\right)^2.$$

- The orthogonality condition

$$\mathbf{F}_N^T \mathbf{\Psi}_N := \sum_{\mathbf{t} \in \mathcal{T}} \mathbf{f}(\mathbf{t}) \psi(\mathbf{t}) = \mathbf{0}$$

  where $\mathbf{F}_N = (\mathbf{f}(\mathbf{t}_1), \cdots, \mathbf{f}(\mathbf{t}_N))^T$ and $\mathbf{\Psi}_N = (\psi_1, \cdots, \psi_N)^T$ with $\psi_i = \psi(\mathbf{t}_i)$.

- Let $\mathbf{\Psi}_N$ vary over a set quantifying the model uncertainty:

$$\mathbf{\Psi} = \{\mathbf{\Psi}_N : \mathbf{F}_N^T \mathbf{\Psi}_N = 0, \ \|\mathbf{\Psi}_N\| \le \tau^2 \},$$

  where $\|\cdot\|$ is the Euclidean norm.

The loss Function

- A nature loss function $\mathcal{L}_0$ is the relative conditional mean squared prediction error (MSPE), averaged over locations in $\mathcal{T}\backslash\mathcal{S}$ at which observations are not obtained:

$$\mathcal{L}_0\left(\xi|\mathbf{\Psi}_N,\theta,\mathbf{d}\right)=\frac{1}{N-n}\sum_{\mathbf{t}\in\mathcal{T}\backslash\mathcal{S}}\frac{E_{\mathbf{y}_n|\mathbf{d},\mathbf{\Psi}_N,\theta}\left(z(\mathbf{t})-\hat{z}_n(\mathbf{t})\right)^2}{z^2(\mathbf{t})}$$

where ξ is the $N\times 1$ 'design' vector, with elements $\xi_i=I\left(\mathbf{t}_i\in\mathcal{S}\right)$.

- A robust design ξ^* optimizes the chosen loss function $\mathcal{L}_0$ in the face of uncertainties.
    - This loss will be averaged, with respect to a 'prior' distribution on $\mathbf{d}$, as a means of relaxing (**U1**).
    - The 'averaged' loss is then maximized over $\mathbf{\Psi}$ to handle (**U2**).

- Upon taking an expectation with respect to $\mathbf{d}$, the loss becomes

$$\mathcal{L}_0\left(\xi|\boldsymbol{\Psi}_N,\theta\right)=\frac{1}{N-n}\sum_{\mathbf{t}\in\mathcal{T}\setminus\mathcal{S}}E_{\mathbf{d}|\boldsymbol{\Psi}_\mathbf{N},\theta}\left(\frac{E_{\mathbf{y}_n|\mathbf{d},\boldsymbol{\Psi}_N,\theta}\left(z(\mathbf{t})-\hat{z}_n(\mathbf{t})\right)^2}{z^2(\mathbf{t})}\right).$$

- It is difficult to maximize $\mathcal{L}_0\left(\xi|\boldsymbol{\Psi}_N,\theta\right)$ with respect to $\boldsymbol{\Psi}_N$.

Expansion of loss Function

- The increasing domain asymptotic framework is an asymptotic framework that the domain is expanding as the number of observations increases.
- The loss function $\mathcal{L}_0\left(\xi|\mathbf{\Psi}_N, \theta\right)$ is expanded up to and including terms that are $O(n^{-1})$ under the increasing domain asymptotic framework.

Expansion of loss Function

Theorem

Apart from terms which are $o(n^{-1})$, the loss function under consideration becomes

$$\mathcal{L}_0\left(\xi|\boldsymbol{\Psi}_N,\theta\right)=\frac{1}{N-n}\sum_{t\in\mathcal{T}\backslash\mathcal{S}}\left[\boldsymbol{\Psi}_N^T\mathbf{A}_{\mathbf{t}\xi\theta}\boldsymbol{\Psi}_N\frac{1}{n}+2\mathbf{b}_{\mathbf{t}\xi\theta}^T\boldsymbol{\Psi}_N\frac{1}{\sqrt{n}}+\left(c_{1\mathbf{t}\xi\theta}+c_{2\mathbf{t}\xi\theta}\frac{\omega_{\mathbf{d}}^2}{n}\right)\right],$$

where

$$\begin{aligned}
c_{1\mathbf{t}\xi\theta} &= E_{\mathbf{y}_n|\theta}\left(1-F_{\mathbf{t}}(\mathbf{0},0)\right)^2, \\
c_{2\mathbf{t}\xi\theta} &= E_{\mathbf{y}_n|\theta}\left[\left(D_{\mathbf{d}}^1F_{\mathbf{t}}(\mathbf{0},0)\right)^2-D_{\mathbf{d}}^2F_{\mathbf{t}}(\mathbf{0},0)+F_{\mathbf{t}}(\mathbf{0},0)D_{\mathbf{d}}^2F_{\mathbf{t}}(\mathbf{0},0)\right], \\
\mathbf{b}_{\mathbf{t}\xi\theta}^T &= E_{\mathbf{y}_n|\theta}\left[\left(F_{\mathbf{t}}(\mathbf{0},0)-1\right)D_{\boldsymbol{\Psi}_N}^1F_{\mathbf{t}}(\mathbf{0},0)\right], \\
\mathbf{A}_{\mathbf{t}\xi\theta} &= E_{\mathbf{y}_n|\theta}\left[\begin{array}{c}-D_{\boldsymbol{\Psi}_N}^2F_{\mathbf{t}}(\mathbf{0},0)+\left(D_{\boldsymbol{\Psi}_N}^1F_{\mathbf{t}}(\mathbf{0},0)\right)^TD_{\boldsymbol{\Psi}_N}^1F_{\mathbf{t}}(\mathbf{0},0)\\+\left(F_{\mathbf{t}}(\mathbf{0},0)D_{\boldsymbol{\Psi}_N}^2F_{\mathbf{t}}(\mathbf{0},0)\right)\end{array}\right].
\end{aligned}$$

Maximization of the loss function over Ψ

Proposition

(Sorensen 1982) The solution $\mathbf{v}^*_{\xi,\theta}$ to the optimization problem is the solution of

$$(\lambda_{\xi,\theta}\mathbf{I}_{N\times N} - \mathbf{A}_{\xi,\theta})\mathbf{v}^*_{\xi,\theta} = \mathbf{b}_{\xi,\theta},$$

and the maximum loss is

$$\mathcal{L}_0\left(\xi|\mathbf{v}^*_{\xi,\theta},\theta\right) = \frac{1}{N-n}\left(\mathbf{v}^{*T}_{\xi,\theta}\mathbf{A}_{\xi,\theta}\mathbf{v}^*_{\xi,\theta} + \mathbf{2b}^T_{\xi,\theta}\mathbf{v}^*_{\xi,\theta} + c_{\xi,\theta}\right), \quad (1)$$

where $\lambda_{\xi,\theta}$ is chosen such that $\lambda_{\xi,\theta}(\left\|\mathbf{v}^*_{\xi,\theta}\right\| - \tau) = 0$ and $\lambda_{\xi,\theta}\mathbf{I}_{N\times N} - \mathbf{A}_{\xi,\theta}$ is positive semi-definite.

- A problem is that this loss depends on the unknown parameters θ. There are various methods of handling this problem.
  - constructing a 'locally optimal' design – one that is optimal only at a particular value of the parameter.
  - To allow for uncertainty about the parameter values, one might first maximize the loss function over a neighbourhood of a local parameter and then minimize the maximized loss function over the class of designs.
  - Bayesian methods are also applicable in eliminating the parameters from the loss function.
- These three methods allow for static, i.e. non-sequential, design construction.

- A method is sequential design.
  - Estimates are computed using the available data and subsequent observations are made at new design points minimizing the loss function, evaluated at the current estimates.

Sequential Robust Optimal Design

- Sequential design:
  - Step 1: choose an initial design ξ_{n_0}.
    For $m = 0, 1, ...$ until an n-point design ξ_n is obtained carry out steps 2-5.
  - Step 2: make observation at the sampled locations of the current design $\xi_m = \{\mathbf{t}_{s1}, ..., \mathbf{t}_{sm}\}$.
  - Step 3: the regression parameters that are required in the evaluation of the loss are replaced by GLS estimation $\hat{\theta}_m$.
  - Step 4: substitute $\widehat{\theta}_m$ into the loss function and obtain $\mathcal{L}_{\max}(\xi_m|\widehat{\theta}_m)$ by maximizing the loss function over the set $\mathbf{\Psi}$.
  - Step 5: make the next observation at

$$\mathbf{t}_{new} = \arg\min_{\mathbf{t}\in\mathcal{T}} \mathcal{L}_{\max}(\{\mathbf{t}_{s1}, ..., \mathbf{t}_{sm}, \mathbf{t}\} \,|\hat{\theta}).$$

Robust design vs. maximin space-filling design

- The true model for $Y(\mathbf{t})$ is

$$y = \mu(\mathbf{t}) + \varepsilon(\mathbf{t})$$

where

$$\mu(\mathbf{t}) = 1 + t_1 + t_2 + \frac{\psi(\mathbf{t})}{7} + \delta(\mathbf{t}),$$

the stochastic error $\delta(\mathbf{t})$ has correlation function

$$corr(\mathbf{t}, \mathbf{t}') = exp\{-0.5||\mathbf{t} - \mathbf{t}'||^2\},$$

$Var[\delta(\mathbf{t})] = 1$ and $Var[\varepsilon(\mathbf{t})] = 0.01$.

- The threshold probability of interest is $P(\mu(\mathbf{t}) > 1)$.

Robust design vs. maximin space-filling design

- The fitted model is

$$y = f^T(\mathbf{t})\boldsymbol{\theta} + \delta(\mathbf{t}) + \varepsilon(\mathbf{t})$$

  with $f^T(\mathbf{t}) = (1, t_1, t_2)$ and $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2)^T$. The nominal covariance matrix is correct.
- The design space $\mathcal{T}$ is a grid of $N = 25$ points spanning $[0, 1] \times [0, 1]$.
- A 3-point initial design was selected at the beginning such that the design points are spread out across the whole grid.
- The sequential procedure is applied to construct a 7-point sequential robust design.

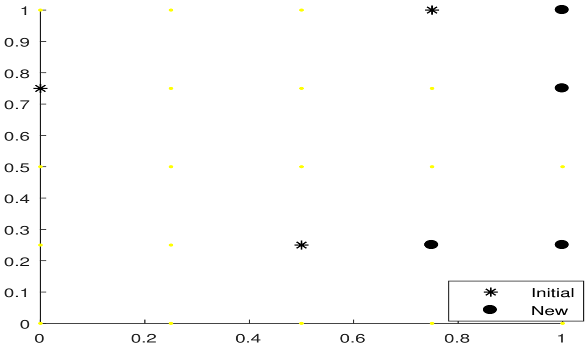

Figure: With 3 initial design points (denoted by asterisks) a robust design (denoted by filled circles) is obtained among the remaining 22 locations.

Robust design vs. maximin space-filling design

- A 7-point maximin space-filling design is selected.

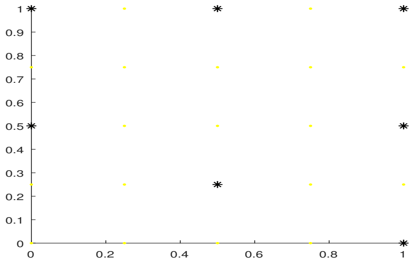

Figure: A 7-point maximin space-filling design is shown with design points denoted by asterisks.

Robust design vs. maximin space-filling design

- To compare performance of the maximin space-filling design and the robust design, the losses for the prediction of the threshold probability were found and summarized in Table 1 when $\tau^2 = 0.25, 0.5, 0.8, 1$.

Design	τ^2 = 0.25	0.5	0.8	1
maximin space-filling	0.3984	0.4130	0.4262	0.4337
robust optimal	0.3607	0.3896	0.4157	0.4304

Table: Losses for maximin space-filling designs and robust designs.

Conclusion

- A method of constructing robust optimal designs for the estimation of threshold probabilities of a stochastic process is proposed.
- Robust optimal designs perform better than the maximin space-filling designs.
- The method is applied for coal-ash data (Hu 2017).

References

- Cressie, N. (1993), *Statistics for Spatial Data*, New York: Wiley.
- Sorensen, D. C. (1982), "Newton's Method with a Model Trust Region Modification," *SIAM Journal on Numerical Analysis*, 19, 409-426.
- Santner, T. J., Williams, B. J., and Notz, W. I. (2003), *The Design and Analysis of Computer Experiments*, Springer Verlag: New York.
- Wiens, D. (2005a), "Robustness in spatial studies I: minimax prediction," *Environmetrics*, 16, 191-203.
- Wiens, D. (2005b), "Robustness in spatial studies II: minimax design," *Environmetrics*, 16, 205-217.
- Wiens, D. and Zhou, J. (2008), "Robust estimators and designs for field experiments," *Journal of Statistical Planning and Inference*, 138, 93-104.

Thank you!